Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 48 (2015) 353 – 359

Procedia Computer Science

International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2015)

Conference Organized by Interscience Institute of Management and Technology,

Bhubaneswar, Odisha, India

Quarter-Sphere Support Vector Machine for Fraud Detection in Mobile Telecommunication Networks

Sharmila Subudhi[a]*, Suvasini Panigrahi[b]

[a]*Ph.D. Scholar, Dept.of CSE, Veer Surendra Sai University of Technology, Burla-768018, India*
[b]*Asst.Professor, Dept. of CSE, Veer Surendra Sai University of Technology, Burla-768018, India*

Abstract

This paper addresses the problem of finding out fraudulent calls of mobile phone users by comparing the most recent call patterns with their past usage patterns. We have modeled the user's profile based on the most relevant fraud detection features like call duration, call type, call frequency along with location and time data. The Reality Mining dataset has been used for testing the efficiency of the proposed methodology. In this work, we discriminate the malicious behavior of users from the normal behavior by training on Support Vector Machine (SVM) classifier. An anomaly is detected when the current pattern of a user (subject) does not match with any of the individual's normal patterns. We have also focused on the improvement of the classifier by applying the concept of Quarter-Sphere SVM. The Quarter Sphere-SVM is a formulation of One-Class SVM, supported by Support Vector Data Description which helps the SVM in unsupervised learning. Our experiments show promising results in terms of detecting fraudulent calls without raising too many false alarms.

Peer-review under responsibility of scientific committee of International Conference on Computer, Communication and Convergence (ICCC 2015)

Keywords: Mobile communication networks; Fraud detection; Support Vector Machine; One Class-SVM; Quarter-Sphere SVM

* Corresponding author. Tel.: +0-787-378-3109; fax: +0-787-378-3109.
*E-mail address:*sharmilasubudhi1@gmail.com

Peer-review under responsibility of scientific committee of International Conference on Computer, Communication and Convergence (ICCC 2015)
doi:10.1016/j.procs.2015.04.193

1. Introduction

The telecommunications industry has emerged substantially in the last few years with the increasing use of mobile phones. Mobile phone fraud is also set to rise with the advancement of mobile phone technology. Mobile phone fraud occurs whenever a fraudster uses deceitful means for availing mobile phone services free of charge or at a reduced rate. This problem prevails worldwide and is the reason of annual revenue losses for many companies. According to a survey report[1] published by one of the Big Four auditor company KPMG on 2012, the global telecom industry suffered losses of around $40 billion which is roughly 2% of the global telecom revenue. Another report[2] published by Communication Fraud Control Association (CFCA) states that the annual global telecom fraud is in between $54 billion and $60 billion. In addition to the financial losses, fraud is the reason of loss of service, loss of reputation and loss of subscriber's confidence[3]. Thus, mobile phone fraud detection problem needs to be addressed in the best possible manner.

The most prevalent type of telecommunication fraud is the superimposed fraud which is usually detected through the presence of fraudulent calls on the bill. The call record comprises of the mixture of some genuine activity along with some malicious activity of a genuine subscriber's account. The illegitimate activity is done by the fraudster other than the account holder. This type of fraud remains undetected for a long time as the number of fraudulent calls is comparatively small according to the overall call volume[4]. For this reason, we focus on identifying this type of fraud.

In this paper, we have proposed a mobile phone fraud detection system (FDS) that employs the application of Support Vector Machine (SVM)[5] and One Class-SVM (OC-SVM) formulated in Quarter Sphere-SVM (QS-SVM) for detection of fraudulent behavior of mobile phone subscribers. We have experimented by using different types of SVMs in the proposed model such as normal SVM and QS-SVM[6] for comparative analysis of the system. To the best of our knowledge, this is the first ever attempt to develop a mobile phone FDS by using QS-SVM.

The rest of the paper is organized as follows. In Section 2, we have discussed the work done in mobile phone fraud detection. Section 3 focuses on the basic concepts of SVM and QS-SVM. In Section 4, we have explained the components of the proposed FDS along with its working methodology. Section 5 presents the results obtained from experimentation. Finally, in Section 6, we have concluded the paper by providing some directions for further enhancements of the model.

2. Related Work

With accordance to the different approaches used by the authors in the literature, some published work is reviewed. Rule-based system[7] has been proposed which uses supervised neural networks for detection of fraud based on the total count of the call duration in a day. A combination of supervised feed-forward neural network, Gaussian mixture model and Bayesian network[8] has been used to detect fraud in communication networks. A fraud detection technique based on the behavioral profiling[9] of the subscribers has been introduced where unsupervised neural network is used for the classification.

Moreover, a fusion of Dempster-Shafer theory and Bayesian inferencing[10] has been proposed for the detection of fraudsters, resulting in better accuracy. The concept of self-organizing map and clustering[11] to detect the fraud has been proposed by monitoring the network traffic. Cloud based intrusion detection system[12] has also been proposed for smart phones which was implemented in Linux kernel. By using the temporal logic of causal knowledge[13] malware activities in a mobile phone are detected. Fraudulent activities in a mobile carrier[14] are done by analyzing the network traffic pattern of users. Fraud detection based on the motion trajectory behavior[15] of a user has been successfully implemented by applying Markov properties.

Although, several techniques have been applied for developing mobile phone FDSs, however, most of the existing systems show a lot of variation in accuracy as they have not considered many relevant call features in building the profile of users. Furthermore, the high rate of false alarms can be addressed by combining observations across time and space dimensions. The main goal of this research work is to address these challenges.

3. Background

Support Vector Machine (SVM) is one of the most sophisticated supervised classification methods in the history of data mining algorithms and has created much speculation in recent years. Further, in case of mobile phone fraud detection, the training of the classifier model needs to be performed regularly in order to adapt to the changing behavior of the mobile phone subscribers over time. This is done by applying the prior knowledge of the fraudulent behavior. Thus, with the help of unsupervised anomaly detection methodologies, we can track the fraudsters. OC-SVM[16] aims to address this issue by helping the SVM in unsupervised learning. It implements a data description boundary around the training data set, known as Support Vector Data Description (SVDD). This boundary can be used to detect genuine or abnormal behavior as it concerns with the characterization of a data set. The main concept of OC-SVM is that data points from input space are first mapped to a higher dimensional feature space using non-linear function. The decision boundary of normal data is then found using SVDD[17], which comprises most of the data points of the feature space in spherical manner. The points lying outside the boundary are classified as anomalous.

The QS-SVM is the formulation of OC-SVM which encapsulates the feature space data points in a hypersphere with a minimal radius. The center of the sphere is fixed. It holds most of the data points present in the feature space. The points falling outside the hypersphere belongs to anomalous data. The optimization problem of the QS-SVM is formalized as follows:

$$\min_{\mathrm{D}\in\mathrm{R},\ \gamma_i\in \mathrm{D}^{\mathrm{b}}} \mathrm{D}^2+\frac{1}{tb}\sum_{i=1}^{b}\gamma_i$$
$$subject\ to\ \ \|\varphi\ (x_i)\|^2\ \leq D^2+\gamma_i\ ,\ \gamma_i\ \geq 0,\ i=1,2,\dots n \tag{1}$$

where D represents the radius, x_i is the input data vector, $\gamma_i = 1,2,\dots n$ is a set of slack variable which allow some x_i to fall outside the hypersphere, b is the number of data vectors and $t \in (0,1)$ is a free parameter which denotes the outlier fraction. The dual formulation on Eq. (1) is given by the following equation:

$$\min_{\alpha\in D^b} -\sum_{i=1}^{b}\alpha_i\, k(x_i,x_i)$$
$$subject\ to\ \sum_{i=1}^{b}\alpha_i=1,\ 0\leq\ \alpha_i\leq\frac{1}{tb},\ i=1,2,\dots.b \tag{2}$$

The data vectors with $\alpha\ >\ 0$ are called *support vectors*, closest to the decision hyperspherical surface, which helps in estimating the accuracy of a SVM classifier. Support vectors with $0\leq\ \alpha_i\leq\frac{1}{tb}$ falling on the sphere are denoted as *marginal support vectors*. Support vectors with $\alpha_i=\frac{1}{tb}$ falling outside the hypersphere are the *non-marginal support vectors* representing the anomalous data.

4. Proposed Fraud Detection Approach

The proposed mobile phone FDS monitors the behavior of a subscriber by comparing the most recent activity patterns with past usage patterns. The working of the FDS is divided into the two main functionalities as discussed below:

a. Profile building of users
b. Mobile phone fraud detection

4.1. Profile building of users

For the building of profiles for different users, we have considered the Reality Mining data set[18]. The dataset is the collection of the traces recorded over a 9-month period for 94 users, holding records of phone call logs, messaging log, location logs such as tower ID, area ID and much more of other information. In the context of the proposed mobile phone FDS, we have extracted the following features for user profile building which is represented as a 6-tuple data set: <*user_id, date_time, call_dur, loc, c_type, c_freq*>.

The feature *user_id* refers to the MAC ID of the mobile device to represent the users uniquely, *date_time* refers to the UNIX timestamp of the call made by the caller, *call_dur* represents the call time duration in seconds. Reality mining data set provides the following information to represent a location cell tower ID and area ID. The cell tower ID gives information associated with the location of the user providing the information for the user's motion. The area ID represents the physical location information like different geographical places. In our approach, we have taken these two features together to represent the *loc*. The *c_type* feature represents the type of the call made by the user, based on following representation: 0 is for local call, 1 for national call and 2 denotes international call and the *c_freq* represents the total number of calls made between two phone numbers in a day.

4.2. Mobile phone fraud detection

The mobile phone fraud detector module deals with the detection of fraudulent calls or activities of user's by employing SVM and QS-SVM. After the training of the SVM classifier is done, a classifier model is generated. The anomalous behavior is detected by applying the test dataset on the model. Thereafter, a comparison is performed in between the test call pattern data vectors and the support vectors present inside the SVM classifier model. An anomaly is detected if there is no match in between them.

4.3. Proposed Fraud Detection Algorithm

The proposed mobile phone fraud detection algorithm is based on the quarter sphere support vector machine.

Input (**I**): *user_id, date_time, c_dur, loc, call_type call_freq*

/*Normalization is performed on the input dataset, **D** = normalized dataset representing input data vectors in [0,1] */
D = normalize (**I**)

/*Dimension reduction of **D** by using PCA, **D_R** = reduced dataset */
D_R = dim_red (**D**)

/* prepare the dataset for giving input in SVM */
$\mathbf{T_r}$ = training dataset (**D_R**) // Training dataset
$\mathbf{T_t}$ = testing dataset (**D_R**) // Testing dataset

/* For training on SVM*/

(C, σ) = crossval_ grid ($\mathbf{T_r}$) //Determine best values of *C* and σ
m = train ($\mathbf{T_r}$, C, σ) //Train the SVM classifier

/* Fraud Detection by applying $\mathbf{T_t}$ on the model */

For a given record $t \in \boldsymbol{T_t}$

if (t match m) then

Output (“Genuine”)

else

Output (“Fraudulent”)

5. Experimental Results

The efficiency of our fraud detection approach is demonstrated by testing it with Reality mining data set. We have first normalized (scaled) the data in the range [0, 1] as SVM only requires a set of real number vectors i.e. either class 0 or class 1. After scaling, dimensionality reduction is performed to obtain a new set of low dimensional data by using Principal Component Analysis[19] (PCA) technique. For the preparation of the training and test dataset, we have used split percentage (split %) method. We have considered 70% split so that the training set is composed of 70% of the dataset and the test dataset comprises of the rest 30%. The training dataset is used to train the SVM classifier and test dataset is used to measure the efficacy of the classifier. After the necessary preparation of dataset is done, a 10-fold cross-validation and grid search method has been used together on the training dataset to find out the best regularization parameters C and σ. These two methods randomly divide the training dataset into 10 equal size folds and find the best combination of C and σ from those folds. These parameters are then used to train the whole classifier model. For our approach, the SVM yields the lowest error rate at the combination values of $C = 8$ and $\sigma = 0.5$. After the classifier model is generated, the test set is used to find out the performance of the proposed FDS in terms of accuracy, true positive rate (TPR) and false positive rate (FPR).

$$Accuracy = \frac{TP+TN}{P+N} \tag{3}$$

$$TPR = \frac{TP}{P} \tag{4}$$

$$FPR = 1 - \frac{TN}{N} \tag{5}$$

True positive (TP) refers to the number of fraudulent samples that were correctly classified by the classifier model. False positive (FP) is the number of genuine samples that were incorrectly labeled by the classifier model as intrusive. We have considered P as the total number of positive samples and N as the total number of the negative samples.

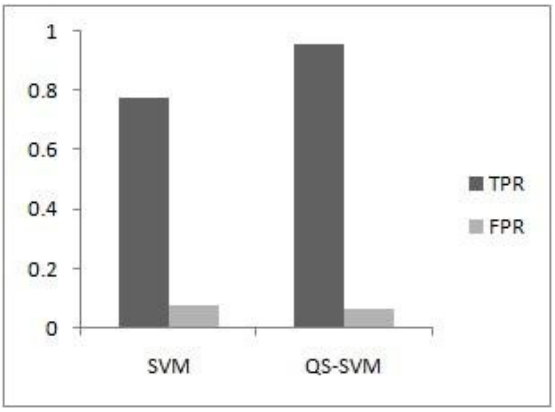

Fig.1. TPR and FPR of SVM and QS-SVM

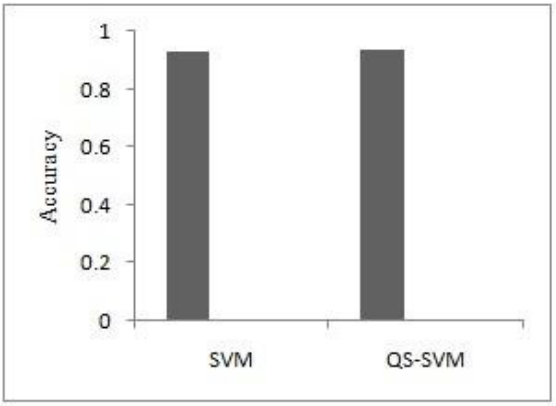

Fig.2. Accuracy of SVM and QS-SVM

Fig. 1 represents the analysis of TPR and FPR values for our proposed model using different SVMs. The result clearly depicts that by using QS-SVM the detection rate increases substantially compared to the usage of SVM classifier. Moreover, the QS-SVM minimizes the false alarm rate effectively and thus raises few alerts as compared

to the normal SVM. In Fig. 1 it is clearly seen that by using QS-SVM, the FDS gives better accuracy. Fig. 3 shows similar comparative analysis using the total execution time of the proposed system using SVM and QS-SVM. It is observed that, the execution time is less by applying QS-SVM which in turn increases the efficiency of the FDS.

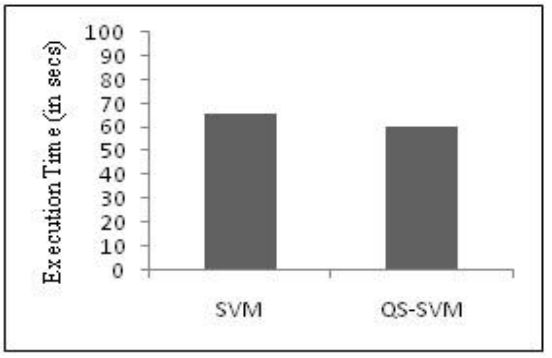

Fig.3. Total Execution Time of SVM and QS-SVM

6. Conclusions

In this paper, we have introduced a novel approach for mobile phone fraud detection by unsupervised learning using One-Class SVM formulated in Quarter Sphere-SVM. The Reality mining data set has been used for evaluating the performance of our proposed approach. Our experimental results reflect the capability of QS-SVM in successfully detecting the fraudulent behavior of mobile phone subscribers while keeping the false alarm rate within a reasonable limit. Based on the results, it can be concluded that the use of QS-SVM addresses this type of real world problems effectively. From the comparative analysis done by using SVM and QS-SVM, it is evident that the use of QS-SVM yields better result than the normal SVM. Although our present work focuses on fraud detection in mobile telecommunication networks, we strongly believe that by employing the application-specific changes, the present approach can be used efficiently for detecting intrusive activities in other applications and generic databases as well.

ACKNOWLEDGEMENT

The authors are highly grateful to the Department of Computer Science and Engineering, Veer Surendra Sai University of Technology (Formerly UCE Burla), Burla, Sambalpur, India for providing the required amenities and support for making this investigation successful.

References

1. Prasad S, *Telecom fraud cost $40b globally*; 2012. http://www.zdnet.com/telecom-frauds-cost-40b-globally-7000008466.
2. Communications Fraud Control Association, http://www.cfca.org.
3. Hoath P. *Telecoms Fraud, The Gory Details*. Computer Fraud & Security; 1999. p. 10–14.
4. Cox KC, Erick SG, Wills GJ. *Visual data mining: Recognizing telephone calling fraud*. Data Mining and Knowledge Discovery: vol.1; 1997. p. 225-231.
5. Cortes C, Vapnik VN. *Support vector networks*. Machine Learning, vol. 20; 1995. p. 273–297.
6. Laskov P, Schafer C, Kotenko I. *Intrusion Detection in Unlabeled Data with Quarter Sphere Support Vector Machines*. In Detection of Intrusions and Malware & Vulnerability Assessment, Dortmund; 2004.
7. Moreau Y, Vandewalle J. *Detection of Mobile Phone Fraud using Supervised Neural Networks: A First Prototype*. Proceedings of the International Conference on Artificial Neural Networks; 1997.
8. Taniguchi M, Haft M, Hollmen J, Tresp V. *Fraud Detection in Communication Networks using Neural and Probabilistic methods*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing; 1998. p. 1241–1244.
9. Burge P, Shawe-Taylor J. *An Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users for Use in Fraud Detectio*n. Journal of Parallel and Distributed Computing; 2001. p. 915–925.
10. Panigrahi S, Kundu A, Sural S, Majumdar AK. *Use of Dempster–Shafer theory and Bayesian inferencing for fraud detection in mobile communication network*. Proceedings of the 12th Australasian Conference on Information Security and Privacy (ACISP). Lecture notes in computer science: vol. 4586; 2007. p. 446–460.

11. Kumpulainen P, Hatonen K. *Anomaly detection algorithm test bench for mobile network management.* Tampere University of Technology; 2008.
12. Houmansadr A, Zonouz SA, Berthier R. *A Cloud-based Intrusion Detection and Response System for Mobile Phones.* IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshop; 2011. p. 31-32.
13. Chaugule A, Xu Z, Zhu S. *A Specification Based Intrusion Detection Framework For Mobile Phones*. Applied Cryptography and Network Security. Lecture Notes in Computer science: vol 6715; 2011. p. 19-37.
14. Jiang N, JinY, Skudlark A, Hsu W, Jacobson G, Prakasam S, Zhang Z. *Isolating and Analyzing Fraud Activities in a Large Cellular Network via Voice Call Graph Analysis*. Proceedings of 10th international conference on mobile systems, applications and services. MobiSys'12; 2012. p. 253-266.
15. Yazji S, Scheuermann P, Dick RP, Trajcevski G, Jin R. *Efficient location aware intrusion detection to protect mobile devices*. Pervasive Ubiquitous Computing. Springer: vol-18; 2014. p. 143–162.
16. Wang D, Yeung DS, Tsang ECC. *Structured one-class classification*. IEEE Transaction on Systems, Man, and Cybernetics. Part B. Cybernetics: vol. 36. no. 6; 2006. p. 1283–1295.
17. Tax DMJ, Duin RPW. *Support vector data description*. Machine. Learning: vol. 54. no. 1; 2004. p. 45–66.
18. Eagle N, Pentland A. *Reality mining: sensing complex social systems*. Personal and Ubiquitous Computing: vol. 10. no. 4. Springer-Verlag; 2006. p. 255–268.
19. Jolliffe IT. *Principal Component Analysis*. 2nd ed. Springer Series in Statistics; 2002.